International Journal of Multidisciplinary Research and Technology
ISSN 2582-7359, Peer Reviewed Journal, Impact Factor 6.325
www.ijmrtjournal.com

A COMPREHENSIVE REVIEW ON EXPLAINABLE AI MODELS FOR HEALTHCARE DIAGNOSIS: BRIDGING ACCURACY AND INTERPRETABILITY

[1]Kamal Khokhar, [2]Dr. Sandeep Kumar Tiwari

[1]MTech Student, [2]Professor, Department of Computer Science & Engineering
Vikrant University, Gwalior MP

Abstract

Artificial Intelligence (AI) has revolutionized the healthcare diagnosis process since now it is possible to diagnose the diseases, predict the risk, and assist in clinical decisions with high accuracy. However, strategies to utilize complex black-box models within critical healthcare systems are hampered by the absence of transparency and accountability as well as interpretability of the complicated black-box models. Explainable Artificial Intelligence (XAI) is now being proposed as a necessary solution to this gap to enhance the transparency, reliability, and clinical usefulness of AI-based diagnostic systems. Some of the XAI methods discussed in this literature review are SHAP, LIME, and Grad-Cam, as well as intrinsically interpretable models, such as decision trees, fuzzy logic systems, and Bayesian networks. The paper entails a comparison analysis of these approaches and describes the trade-offs among the other healthcare areas, such as radiology, genomics, EHR-based prediction, and critical care. It also addresses hybrid XAI systems, which deploy deep learning on explainable systems in the quest of the high performance and trustful extraction of decisions. As noted in the review, XAI is critical to support clinician trust, ethical and legal standards, and safe AI adoption in clinical processes. The future trends in the research are addressed to enable the development of robust, interpretive, and clinically useful AI systems to transform healthcare diagnosis.

Keywords: Artificial Intelligence (AI), Explainable AI (XAI), Healthcare Diagnosis, Interpretability, Transparency, SHAP

Introduction

In recent years, Artificial Intelligence (AI) has become a revolution within the healthcare industry as it enables the use of complex disease diagnosis, treatment plan, and patient running [1]. Machine learning (ML) and deep learning (DL) have been shown to perform exceptionally well in tasks such as the medical analysis of images and predictive modelling of patient outcomes, sometimes surpassing human activity, and can be applied to provide more effective care to patients, reduce errors in diagnosis, and automate health care delivery [2]. Regardless of these successes, the majority of AI systems are considered black boxes, which produce predictions without explaining the mechanism and this factor is associated with the issue of trust, accountability, and ethical decision-making in the clinical practice. Clinicians and healthcare stakeholders require interpretable insights to defend the decisions, assure patients of their safety, and make sure that they adhere to the regulatory provisions. Explainable Artificial Intelligence (XAI) has risen up to these challenges by coming up with methods through which AI models can be transparent and understandable without having to significantly reduce their performance [3]. XAI increases trust and assists in making decisions by offering explanations that are human-understandable, which is why it can help to integrate AI into standard healthcare practice. This review offers an analytical review of the XAI methods in healthcare diagnosis that include both intrinsically interpretable models that include decision trees and rule-based systems, post-hoc explanation methods such as SHAP, LIME, and Grad-CAM, and how the trade-offs between accuracy and interpretability can be addressed, how XAI methods can be practically implemented in clinical settings, and what future research needs to be done to make AI applications both trustworthy and clinically useful.

Table 1: Summary of Key Studies on Explainable AI (XAI) in healthcare

Author(s) & Year	Title / Source	Focus Area	XAI Methods Discussed	Key Contributions / Findings
Ozdemir & Fatunmbi (2024) [4]	*Explainable AI (XAI) in healthcare: Bridging the gap between accuracy*	Healthcare XAI, model transparency	SHAP, LIME, interpretable ML models	Highlights the importance of balancing accuracy & interpretability;

	and interpretability – Journal of Science, Technology & Engineering Research			discusses XAI adoption barriers and proposes frameworks for clinical deployment.
Tuan (2024) [5]	*Bridging the gap between black box AI and clinical practice*	Clinical integration of XAI	Broad XAI frameworks	Focuses on trust, ethics, and personalized diagnostics; emphasizes need for human-centered explanation systems.
Ennab & Mcheick (2024) [6]	*Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review* – Frontiers in Robotics and AI	Challenges & future directions in healthcare XAI	SHAP, LIME, Grad-CAM, rule-based models	Identifies limitations in current XAI; provides roadmap for robust, reliable, and ethical XAI-based medical systems.
Narkhede (2024) [7]	*Comparative evaluation of post-hoc explainability methods in AI: LIME, SHAP, and Grad-CAM* – IEEE ICSES	Comparative assessment of XAI tools	LIME, SHAP, Grad-CAM	Compares effectiveness, stability, and clinical usability; shows domain-dependent performance variations.
Hooshyar & Yang (2024) [8]	*Problems with SHAP and LIME in interpretable AI for education* – IEEE Access	Post-hoc explainability challenges	SHAP, LIME, neural-symbolic extraction	Demonstrates limitations of SHAP/LIME such as instability; proposes neural-symbolic rule extraction as a more interpretable alternative.

1. Background on AI in healthcare diagnosis

Artificial Intelligence (AI) has quickly revolutionized the process of healthcare diagnosis because machines can process complicated medical records, see patterns, and assist with clinical judgments at a very high level of precision [9]. To find diseases and predict outcomes and recommend individualized treatment, machine learning (ML) and deep learning (DL) algorithms are actively applied to process large amounts of patient data, including electronic health records, medical imagery, and genomic data. The AIs have shown remarkable ability to identify such conditions as cancer, cardiovascular diseases, and nervous disorders and are often more efficient than the conventional ones and assist clinicians to get rid of their mistakes and improve their efficiency. The introduction of AI into the sphere of healthcare diagnosis has not only enhanced the accuracy of its results and speed but has also enabled to spot the diseases at their inception, and avert them, which contributed to the overall increase in patient outcomes. However, with the increase in complexity of AI models, and deep learning networks, in particular, the problem of interpretability, trust, and ethical use has also become apparent, thus, necessitating the exploration of the possibility of transparent and explainable approaches that would enable safe and trustworthy clinical application.

2. Challenges of black-box AI models in critical healthcare decisions.

Black-box AI models in healthcare are highly accurate in their predictions but non-transparent, and clinicians find it difficult to come up with a rationale behind the choice. This ambiguity prompts conflict regarding trust, responsibility, application of AI ethically and safely, and its integration within the key clinical practices [10].

- **Lack of Transparency:** Complex deep learning systems are black-box AI models who make predictions without giving a clear explanation of the decision making processes. Such a lack of transparency does not allow clinicians to know the rationale behind a diagnosis or a recommendation and it is also hard to determine the trustworthiness of the AI system.

- **Limited Trust and Adoption by Clinicians:** Healthcare professionals become reluctant to base their decisions on AI-generated results when the decision-making process cannot be interpreted. The lack of transparency regarding the way conclusions are obtained can dishearten clinicians, which will impede the implementation of AI tools into clinical practice.
- **Accountability and Legal Concerns:** Black-box models complicate the identification of responsibility in the event of misdiagnosis, or a wrong treatment. The fact that internal logic of the AI system is not transparent makes assigning accountability to the system, the developers, and clinicians a complicated issue that presents legal and ethical concerns.
- **Ethical and Bias Issues:** Black-box models have inadvertent risks of bias in training data and result in unfair or discriminate results. Unless interpretable, these biases are hard to identify or rectify, a fact that may jeopardise ethical practices and patient safety.
- **Difficulty in Clinical Validation and Integration:** Clinical use of AI predictions requires validation of the predictions by established medical knowledge. The black-box models complicate this process and make it hard to have AI-regulation approved and constrain AI integration to the present healthcare workflows. The interpretability is also not applicable to effective communication of AI-driven decisions to patients by clinicians.

Overview of AI in Healthcare Diagnosis

Artificial intelligence (AI) has caused a substantial change in healthcare diagnosis with the ability to analyze a large volume of medical data quickly and accurately. The broad application of AI techniques, especially machine learning (ML) and deep learning (DL), to process electronic health records, medical imaging, genomic data, and other clinical data is in the detection of diseases, forecasting patient outcomes, and aiding in the treatment plan [11]. Its common uses are in the early diagnosis of cancers with the help of imaging, the detection of cardiovascular and neurological diseases, and disease prognosis. Diagnostic tools based on AI increase accuracy and minimize the human aspect, as well as efficiency in clinical decision-making. Different AI models such as traditional algorithms such as decision trees and support vectors machines, to more complex neural networks, have been effectively utilized in different fields of healthcare. These models have shown remarkable performance, but due to their complexity, such interpretability is usually restricted, thus the need to have explainable AI solutions that will allow transparency, trust, and safe adoption into clinical practice.

1. Common AI techniques used in healthcare

Artificial Intelligence (AI) is a field that utilizes numerous methods to assist in healthcare diagnostics, treatment plans, and patient management through the analysis of multidimensional medical data. In general, these methods can be divided into two, namely, Machine Learning (ML) and Deep Learning (DL). The use of ML algorithms is especially useful with organized data, including electronic health records, lab findings, demographics, etc., which allows the implementation of activities, like disease forecasting, risk evaluation, and patient categorization. DL techniques, in turn, are good at manipulating unstructured data, such as medical images, genomic sequences and time-series data, and can be used to provide high-quality diagnostics and sophisticated pattern recognition. Combined, such AI solutions will offer an effective set of tools to enhance clinical decision-making, increase diagnostic efficiency, and allow customized healthcare solutions.

- **Machine Learning (ML) in healthcare:** Machine Learning (ML) is a type of artificial intelligence, which is popular in the healthcare industry, where the focus is on developing the models that can learn patterns or patterns with the help of structured data to predict or classify input information. The most commonly utilised algorithms to analyse electronic health records, laboratory results and patient demographics are random Forest, Support Vector Machines (SVM) and XGBoost [12]. These models can determine the patterns of diseases, forecast the outcome of the patients and assist in early diagnosis with high accuracy. ML approaches are useful with structured or tabular data, and provide a sort of interpretation, hence explaining why they would be useful in clinical decision support systems.
- **Deep Learning (DL) in Healthcare:** Deep Learning (DL) is a subfield of AI, meaning an automatic derivation of features in highly unstructured and complex data, such as medical images, time-sequences, and genomes, on the basis of neural networks. Image-based diagnostics, such as MRI, CT scans, and X-ray analysis, are calculated with the help of such methods as Convolutional Neural Networks (CNNs). RNNs are used on sequential data, including patient monitoring or time-series clinical data. With the help of transformers,

more complex healthcare patterns can be recognized in multimodal healthcare data, enhancing predictive outcomes on a variety of medical tasks. DL models can be very precise but are usually black boxes, which emphasizes the role of interpretable AI in order to guarantee a level of interpretability and trust within clinical environments.

Applications of AI in Healthcare

AI is present in diverse applications in healthcare, making medical services more accurate, efficient, and personal.

1. Disease Prediction and Diagnosis: Patient data (electronic health records, lab outcomes and genetic data) is analyzed using AI models to identify patterns and risk factors of various diseases. It assists in the detection and intervention before it is too late and also diagnoses are much more accurate to ensure that the clinicians make correct decisions and the probabilities of error reduction.

2. Imaging Analysis (X-rays, MRIs, CTScans): Medical imaging is appealing to the deep learning techniques, namely CNNs. Artificial intelligence can detect any abnormalities, classify diseases and point to problematic areas with high precision using X-rays, MRIs and CTs. This will assist radiologists in improving the quality of diagnosis made, reduce the workload and also expedite the imaging assessment process.

3. Personalized Treatment Recommendations: AI systems are capable of consolidating various patient data, to propose personalized treatment options, considering medical history, genetic and lifestyle data. This helps in precision medicine in streamlining the choice of therapy, anticipating patient response to therapy and enhancing healthcare outcomes in general.

These applications prove the effectiveness of AI in improving clinical decisions, error reduction, and assisting in the provision of health services that is more effective and patient-centered.

2. Limitations of Black-Box AI Models

1. **Lack of Transparency:** Deep learning systems and other black-box AI models are highly non paralleled, comprising of many layers and interrelated parameters. Although they do make correct forecasts, they cannot be seen and comprehended by human beings in the process of making decisions. This non-disclosure in healthcare is critical as to make an informed decision a clinician must know why a diagnosis or recommendation was made. In the absence of explanations, it is hard to determine how reliable AI predictions may be and rely on the system in high-stakes scenarios like life-threatening diagnosis.
2. **Limited Trust and Adoption:** Health care can lack confidence in opaque AI models, as these models are not transparent. The reluctance of clinicians to incorporate AI systems in their practice might be associated with the inability to read or comprehend the manner of drawing conclusions. Such mistrust is likely to hinder the implementation of potentially useful AI solutions since medical professionals would like to use the solutions offering interpretable results and can help in making evidence-based decisions. It is also less likely to promote confidence in AI-driven systems because it is impossible to explain a decision to other employees or patients.
3. **Difficulty in Accountability:** The issue of accountability is raised when black-box AI systems are involved in clinical practices. When a model makes a wrong diagnosis or treatment suggestion, it may be difficult to decide whose responsibility it is, either the AI developer, healthcare institution, or the clinician. This impossibility to interpret an error is why it becomes hard to identify the cause of the error or take corrective measures and find an appropriate person to blame. This dilemma presents a problem in patient care both legally and ethically.
4. **Ethical and Bias Concerns:** Black-box AI models are trained based on past healthcare data, and they can have certain biases in regard to the demographics, socioeconomic status, or medical practices. These biases can be latent because there were not obvious how the inner functioning of the model works and can result in the unfair treatment or even disadvantageous treatment of some of the groups of patients. Among them, a minority group could be underdiagnosed by the model as an AI due to its underrepresentation in the training data. This is not easy because it is necessary to address such ethical and bias-related concerns without interpretable AI methods, which can facilitate the possibility to analyze the decision pathways.
5. **Challenges in Clinical Validation:** To ensure safety and usefulness of AI technology in medical care, it is necessary to confirm its conclusions by the existing medical information. This is not a simple task since the black-box models are not simple. There is no straightforward way that clinicians and regulators can know

whether the reasoning of the model is in line with clinical standards and scientific evidence. This drawback makes regulation a challenging process, the integration into the current workflows of healthcare more difficult, and the possibility to explain what is being done to the patients to make an informed decision and trust the medical intervention.

Explainable AI (Xai) in Healthcare

Artificial Intelligence (AI) has shown impressive results in the medical field, yet its usefulness is usually restricted by the inability to provide transparency and interpretability of intricate models [13]. Explainable Artificial Intelligence (XAI) works on this issue by creating system and technique to render AI systems interpretable and comprehensible by people, specifically clinicians and health care stakeholders. XAI is significant in healthcare because it can be used to build trust, assure accountability, support regulatory compliance, and provide a healthy clinical decision-making process. XAI assists clinicians to validate AI predictions, detect possible mistakes, and discuss decisions writtenly with patients by explaining AI predictions.

Key Concepts

- Interpretability: Refers to how well humans are able to comprehend the inner logic or logic of an AI model. High interpretability enables clinicians to understand the calculations made by a model to reach its conclusions.
- Transparency: The transparency of an AI system in disclosing its form, information, and algorithm. The transparent models enable the stakeholders to investigate the way predictions are made.
- Trustworthiness: The trust that clinicians and patients will have in the AI systems, which will be based on the reliability, consistency, and explainability of the model outputs.

Types of XAI Approaches

1. Post-hoc Interpretability: Post-training methods used to interpret predictions of black-box models that are complex and do not alter the model [14]. Common methods include:

- SHAP (SHapley Additive exPlanations): Measures the value of individual feature to the prediction of the model.
- LIME (Local Interpretable Model-Agnostic Explanations): Provides local explanations, through the approximation of black-box predictions by interpretable models.
- Grad-CAM (Gradient-weighted Class Activation Mapping): Identifies areas in medical images on which the model makes decisions, and is commonly applied in imaging diagnostics.

2. Intrinsic Interpretability: Models that are simple in structure and which are inherently transparent and easy to comprehend, e.g:

- Decision Trees: Model decisions and their potential results on a tree diagram.
- Rule-Based Models: Make decisions with the help of if-then rules, with clear explanations of the predictions.

Evaluation Metrics for Explainability

Effectiveness of XAI methods is important in clinical adoption. Some of the common evaluation metrics are included [15]:

- Fidelity: Measure of the goodness of the explanation of the real behavior of the model.
- Comprehensibility: Measures the understanding ability of a human on the explanation.
- Consistency: Tests whether similar explanations are obtained with similar inputs.
- Trust and Usability: Assesses the degree of reliance and use of the explanations by clinicians in decision-making.

Overall, XAI is a concept that fills the gap between the accuracy and interpretability of models, rendering AI systems safer, more reliable, and appropriate to be integrated in the healthcare diagnosis and treatment planning.

AI Models for Healthcare Diagnosis

Explainable Artificial intelligence (XAI) has become a very crucial aspect of the healthcare diagnostic system, where transparency, reliability, and accountability are required of critical targets. Although highly predictive models of AI and deep learning have been well-established to be highly predictive, clinicians tend to hesitate in using them because they are considered black-box [16]. XAI fills this gap as it offers interpretable and reliable explanations to allow medical practitioners to see how a model comes up with a decision. In this section, a review of the key Explainable AI methods applied to medical diagnosis will have been made, their applicability assessed, and the trade-offs between interpretability and prediction ability discussed.

1. Rule-Based and Interpretable Models

Rule-based and naturally interpretable models are constructed in such a way that the processes of making decisions can be easily recognized. These models do not only give black-box neural network predictions, but the networks are explicit and give reasons behind predictions.

Decision Trees

Decision trees are inherently interpretable models, which work in a hierarchical form of simple IF-THEN rules. Every internal node is a decision that was made depending on a certain clinical attribute and every leaf node is the final prediction or classification. The open nature of decision trees contributes to their utility specifically in the healthcare field where clinicians are drawn to models that explicitly explain their diagnostic rationale. They are popular in categorizing diseases according to the symptoms, lab parameters, demographic characteristics, and patient history. Typically, it finds use in diabetes prediction on the basis of glucose level, BMI, and age; heart disease classification by vital signs and blood test outcomes; and cancer stage classification based on imaging characteristics and biochemical markers. The fact that they can demonstrate a clear path of decision enables the medical practitioners to accept and believe the computation diagnostic results.

Fuzzy Logic Models

Fuzzy logic models are created to handle uncertainty and imprecision- aspects that are very rampant in clinical settings of decision-making. Contrary to the binary logic, the fuzzy systems operate with linguistic terms (low, medium, and high) to express the variables in the medical field, enabling the system to be more flexible and closer to human logic. The strategy can be used to interpret patient data at a fine level, especially when clinical values are under ambiguous or overlapping values. Liver diagnosis Fuzzy models have been used to diagnose liver diseases where biological measure is continuously varied, to monitor ICU patients whose symptoms can change quickly, and to formulate symptom-based disease scoring systems that can imitate human judgment. Their flexibility to include indistinct information or missing data render them the most appropriate to medical settings [17].

Bayesian Networks

Bayesian networks are the probabilistic graphical models which represent the conditional dependencies amongst the variables in medicine. They are also effective in making predictions and explanations because they enable clinicians to trace the role of various attributes of patients to a diagnostic outcome. Probabilistic interpretation, ability to provide graph-based explainability, and the ability to incorporate expert knowledge of medicine into the network structure are provided by these models. Bayesian networks find application in healthcare to predict disease progression e.g. cancer metastasis; stroke risk prediction based on variables such as blood pressure, cholesterol levels and smoking history; and in developing diagnostic systems to identify respiratory disorders. Uncertainty and causal relationship are other characteristics of them that make them a key resource to the sophisticated clinical decision-support systems.

Applications and Case Studies

Several real-world healthcare systems use these interpretable models:

- **Mycin** (expert system for infectious disease diagnosis) used rule-based logic.

- **PHES** (Pharmacogenomics Health Engine System) uses Bayesian networks to predict drug response.
- Decision trees and fuzzy systems are widely applied in diabetes risk calculators and neonatal health assessment systems.

2. Post-hoc Explainability Methods

The post-hoc XAI methods are used when a model is already trained. They assist in explaining complicated models of deep learning or ensemble models without making modification in their internal structure.

SHAP (SHapley Additive exPlanations)

SHAP is among the most mathematically rigorous post-hoc explainability procedures applied to the contemporary AI models. It gives contribution scores with each feature in the dataset, which is a clear indication of the extent to which each variable contributed to the ultimate prediction. The approach can be especially useful in the field of healthcare given the fact that it can be used to point out significant risk factors associated with various diseases, e.g., the role of HbA1c as a significant risk factor of diabetes. SHAP also gives clear explanations as to why a patient is considered as high-risk and assists in interpreting feature significance in ICU mortality prediction systems. Although the main advantage is that it has a solid theoretical basis and has a coherent explanation, the significant weakness is that it is quite expensive to compute, particularly when used in deep learning models with millions of features [18].

LIME (Local Interpretable Model-agnostic Explanations)

LIME aims at explaining single predictions, constructed using a simplified model of interpretation, on an example. In healthcare diagnostics, the use of LIME is typical to describe why a tumor can be benign or malignant, why there are misclassification trends in the analysis of ECG and EEG signals, and to interpret predictions of cardiovascular diseases. Its main strength is that it is versatile, it can be implemented to any machine learning model irrespective of how complex it is. It, however, has a strong weakness in form of its instability; local explanations can change across different runs, which is why its reliability is lower than that of more stable frameworks such as SHAP.

Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM is a strong visualization instrument and is highly applied in image-related medical scenarios. It also identifies the local areas of a medical image that give the greatest contribution to the classification result of a model, which facilitates the interpretation of deep learning predictions by clinicians. Grad-CAM is commonly used in the identification of pneumonia or COVID-19 on chest X-rays, identifying tumor areas in MRI and CT images, detecting retinal abnormalities in diabetic retinopathy detection, etc. Its greatest strength is that it presents a visual explanation of its results in easily understandable heatmaps that are intuitive and non-disruptive to clinical diagnostic workflows. Its main weakness though is that it is only applicable to convolutional neural network (CNN) architectures, limiting it to non-CNN models.

3. Hybrid Approaches

Hybrid XAI methods are methods that integrate predictive ability of deep-learning with interpretable ones to deliver accuracy and transparency [19].

Combining Deep Learning with Explainable Techniques

These hybrid systems include:

- Deep neural networks + rule extraction
- CNN models + attention mechanisms
- Ensemble models + SHAP/LIME explanations
- Interpretable surrogate models such as interpretable neural additive models (NAMs)

Benefits

- Has high performance and is interpretable.
- Enables clinicians to test model reasoning.
- Increases the confidence of automated diagnosis.

Examples in Clinical Imaging and Patient Data Analysis

Hybrid models have been successful in:

- **Breast cancer detection:** Neural network classifiers based on Grad-CAM and SHAP.
- **Cardiovascular disease prediction:** Decision-rule extraction of ECG features was added to deep learning models.
- **ICU mortality prediction:** Hybrid models using deep neural networks + Bayesian reasoning.
- **Radiology:** Attention layers in CNN constructing heatmaps to support the choice.

4. Comparative Analysis

Table 2: Accuracy vs. Interpretability in Healthcare AI **[20]**

Aspect	Details
Trade-offs Between Accuracy and Interpretability	- **Interpretable models** (Decision Trees, Bayesian Networks): Greater transparency, which is more comprehensible to clinicians; could be inaccurate with complex datasets. - **Deep learning models**: Excellent predictive power; poor interpretation. - **Hybrid models**: Effort to trade off accuracy and explainability. - **Clinical requirement alignment**: ICU decisions need to be interpretable, whereas high-accuracy deep learning with post-hoc XAI can be used in early screening.
Performance Evaluation Across Healthcare Domains	- **Radiology & Imaging**: Grad-CAM and attention-based hybrid methods are the most effective. - **EHR-based predictions**: Rule-based and SHAP models are more interpretable. - **Genomics & High-dimensional data**: SHAP and LIME were the choice of explaining complex models. - **Critical care**: Bayesian networks that are helpful as a result of uncertainty reasoning. **Key evaluation metrics**: Precision, accuracy, recall, F1-score, interpretability score, clinical expert feedback, reliability and bias detection.

Conclusion

Explainable Artificial Intelligence (XAI) is an essential factor in the development of AI-based healthcare diagnosis and solving the lack of transparency and trust of black-box models. This review has indicated that the deep learning models can be used to achieve remarkable accuracy, but their uninterpretability is a serious issue to their clinical implementation, compliance with ethics, and patient safety. Decision trees, fuzzy logic systems and Bayesian networks present interpretable models that provide transparent decision paths but do not always work with complex high-dimensional data. Post-hoc techniques such as SHAP, LIME, and Grad-CAM offer very valuable devices used to understand complex models, but hybrid methods facilitate trade-offs between predictive performance and explainability in a variety of medical problems. Comparative research points out that XAI methods are not standardized, and there is a need to tailor interpretability to area-specific requirements, i.e. when dealing with high-risk patients, e.g. ICU or emergency. Overall, it is possible to consider XAI as a mediator between technological innovation and clinical trust, where AI systems are trustworthy, understandable, and ethical. It will require further development of hybrid solutions, regulators framework, and patient-centered assessment to introduce XAI to its full

capacity in everyday healthcare delivery and realize its potential benefit of improving diagnostic accuracy, patient outcomes, and healthcare provision.

References

1. Rane, N., Choudhary, S., & Rane, J. (2023). Explainable artificial intelligence (XAI) in healthcare: interpretable models for clinical decision support. Available at SSRN 4637897.
2. Reddy, A. K., Thota, S. K., Saini, V., Chitta, S., & Bojja, S. G. R. (2024). Bridging AI and Human Understanding: Interpretable Deep Learning in Practice. Journal of Informatics Education and Research, 4, 3706.
3. Ennab, M., & Mcheick, H. (2022). Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. Diagnostics, 12(7), 1557.
4. Ozdemir, O., & Fatunmbi, T. O. (2024). Explainable AI (XAI) in healthcare: Bridging the gap between accuracy and interpretability. Journal of Science, Technology and Engineering Research, 2(1), 32-44.
5. Tuan, D. A. (2024). Bridging the gap between black box AI and clinical practice: Advancing explainable AI for trust, ethics, and personalized healthcare diagnostics.
6. Ennab, M., & Mcheick, H. (2024). Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. Frontiers in Robotics and AI, 11, 1444763.
7. Narkhede, J. (2024, October). Comparative evaluation of post-hoc explainability methods in ai: Lime, shap, and grad-cam. In 2024 4th International Conference on Sustainable Expert Systems (ICSES) (pp. 826-830). IEEE.
8. Hooshyar, D., & Yang, Y. (2024). Problems with SHAP and LIME in interpretable AI for education: A comparative study of post-hoc explanations and neural-symbolic rule extraction. IEEE Access.
9. Band, S. S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., ... & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked, 40, 101286.
10. Ennab, M. M. H. (2025). A hybrid convolutional-fuzzy model for interpretable AI in healthcare: improving transparency and accuracy in chronic disease management (Doctoral dissertation, Université du Québec à Chicoutimi).
11. Singhal, A., Pratap, P., Dixit, K. K., & Kathuria, K. (2024, March). Advancements in explainable AI: Bridging the gap between model complexity and interpretability. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 675-680). IEEE.
12. Zeb, S., Nizamullah, F. N. U., Abbasi, N., & Fahad, M. (2024). AI in healthcare: revolutionizing diagnosis and therapy. International Journal of Multidisciplinary Sciences and Arts, 3(3), 118-128.
13. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4).
14. Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G. P., ... & Islam, S. R. (2020). Medical diagnostic systems using artificial intelligence (AI) algorithms: principles and perspectives. Ieee Access, 8, 228049-228069.
15. Wu, S., Wang, J., Guo, Q., Lan, H., Zhang, J., Wang, L., ... & Chen, Y. (2022). Application of artificial intelligence in clinical diagnosis and treatment: an overview of systematic reviews. Intelligent Medicine, 2(02), 88-96.
16. Khanna, N. N., Maindarkar, M. A., Viswanathan, V., Fernandes, J. F. E., Paul, S., Bhagawati, M., ... & Suri, J. S. (2022, December). Economics of artificial intelligence in healthcare: diagnosis vs. treatment. In Healthcare (Vol. 10, No. 12, p. 2493). MDPI.
17. Voutouri, A., Kostina, A., Menelaou, P., Bratsa, M., & Sachmpazidis, S. (2024). Overview of AI and Machine Learning in Healthcare. IEEE Journals & Magazine. Retrieved from IEEE Xplore.
18. Hulsen, T. (2023). Explainable artificial intelligence (XAI): concepts and challenges in healthcare. Ai, 4(3), 652-666.
19. Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of explainable AI techniques in healthcare. Sensors, 23(2), 634.
20. Pawar, U., O'shea, D., Rea, S., & O'reilly, R. (2020, June). Explainable AI in healthcare. In 2020 international conference on cyber situational awareness, data analytics and assessment (CyberSA) (pp. 1-2). IEEE.

Use for Citation: Kamal Khokhar, Dr. Sandeep Kumar Tiwari. (2025). A COMPREHENSIVE REVIEW ON EXPLAINABLE AI MODELS FOR HEALTHCARE DIAGNOSIS: BRIDGING ACCURACY AND INTERPRETABILITY. International Journal of Multidisciplinary Research and Technology, 6(12), 16–24. https://doi.org/10.5281/zenodo.17896981